SPEECH RECOGNITION IN 7 LANGUAGES

Ulla Uebler

Bavarian Research Center for Knowledge Based Systems (FORWISS)
Research Group for Knowledge Processing
Am Weichselgarten 7
D-91058 Erlangen, Germany
e-mail: uebler@forwiss.de

ABSTRACT

In this study we present approaches to multilingual speech recognition. We first define different approaches, namely portation, cross-lingual and simultaneous multilingual speech recognition and present results in these approaches. In recent years we have ported our recognizer to other languages than German. Some experiments presented here show the performance of cross-lingual speech recognition of an untrained language with a recognizer trained with other languages. Our results show that some languages like Italian are per se easier to recognize with any of the recognizers than other languages. The substitution of phones for cross-lingual recognition is an important point and we compared results in cross-lingual recognition for different baseline systems and found that the number of shared acoustic units is very important for the performance.

1. INTRODUCTION

Over the years we have studied speech recognition and speech understanding systems in German, and as more and more multilingual applications are needed, the ISADORA system was also used for multilingual speech recognition [1, 8].
The need for multilingual speech recognition applications has risen for example by the growing internationalism like within the European Community or in telecommunications. Thus, applications are developed for recognition in a new language, for example dictation systems are ported to a new language or information systems are developed for e. g. tourist information at airports and train stations which have to be able to understand a couple of languages.
When developing a recognition system for a new language either exclusively for the new language or for the new language in addition to existing languages, the recognition system optimized for the first language has to be adapted to the characteristics of the new language.
During this process, mainly data like the vocabulary, acoustic parameters, language models, and the dialog structure have to be adapted. Most of these adaptations have already been performed before, e. g. when porting a system to a new domain. One topic is still specific to the portation to a new language: the definition and the use of acoustic units. If the recognizer is completely rebuilt for a new language with training material of that language, the definition of new acoustic units arises from the pronunciation of the words in the vocabulary, but when there is not sufficient training material available for the new language or when two languages are recognized at the same time, the acoustic units of the old and the new language have to be set in relation. This problem and solutions to it will be the central aspect in this contribution.
In the following, we will cluster approaches of multilingual speech recognition in order to provide clear definitions for the different approaches and describe characteristics of these approaches. Then we will shortly describe the available data material for our experiments and present different strategies of phone substitution during the transition of languages. We will present experiments and results for different approaches of multilingual speech recognition and phone substitution techniques.

2. DEFINITIONS

When looking at the approaches made in multilingual speech recognition, we find that they may be clustered into three groups depending on the application goal and available data, namely porting, cross-lingual recognition and simultaneous multilingual speech recognition.
When a speech recognition system developed for one language is used for recognition in another language, we speak of *porting*. This step is similar to that of developing an application in a new domain of the same language. The vocabulary and the acoustic units have to be defined for the new language. Special attention must be paid to characteristics of languages like homophones or compound words and other characteristics affecting the recognition process. For these characteristics, algorithms have to be found that can cope with these new problems. The system is then trained with data of the new language. This approach can be found for example in [2, 3, 11].
Another approach follows the same application goal as the approach above with the only difference, that there is not sufficient training material available in the new language. Thus, for *cross-lingual* recognition methods must be found to use training material of another language for a rough modeling of acoustic parameters and only to perform an adaptation with few data of the goal language. One main problem is to determine identical acoustic units or to model existing acoustic units in a way that with few adaptation data a good recognition can be provided. Approaches of this kind can be found for example in [4, 7].

The third cluster of approaches is that of *simultaneous multilingual recognition* . Applications of this approach allow utterances of different languages at the same time for the same recognition system. There are a two main strategies for this approach: firstly, to perform some kind of language identification and perform then monolingual recognition or to have only one recognizer that distinguishes in some way between the languages. For this latter strategy, identical acoustic units may be used across the languages or completely different acoustic units as well as sets of mono- and multi-lingual acoustic units. Also, for language modeling, it may be determined between multi- and monolingual language modeling, which also means that transitions between languages are allowed or not. Approaches for simultaneous speech recognition can be found for example in [1, 8, 10].

3. DATA BASES

The data used in our experiments result from three projects: the EU project SQEL (Spoken Queries in European languages), the EU project SPEEDATA (Speech Recognition for Data-Entry), and from the BMBF project VERBMOBIL.

The SQEL project covers the languages Slovak, Slovenian and Czech in an information system for train and flight time tables. The SPEEDATA project covers the languages Italian and German, both spoken by dialect and non-natives speakers. The task of the project is the entry of land register data in the bilingual region of South Tyrol in the original language, thus the rate of non-native speech will always be around 50 percent. The VERBMOBIL project deals with date scheduling among humans in Japanese, English and German including automatic translation among the languages.

An overview on the training data used from these projects is given in Table 1. With these data, we cover seven languages (German (G1, G2) , Italian (It), Slovak (Sa), Slovenian (Se), Czech (Cz), Japanese(Jp), and English (En)), while German is covered twice. The German data assigned with G1 result from the SPEEDATA project and contain dialect and non-native speakers whereas the data set G2 from the VERBMOBIL project covers only native German speech.

Language	G1	It	Sa	Se
Data/hours	8.6	7.6	5.1	6.1
Distinct vocabulary	5455	6748	1061	955
	Cz	Jp	En	G2
Data/hours	7.2	27.4	9.6	28.5
Distinct vocabulary	1323	3207	2157	7444

Table 1. Acoustic data for each language

The data consist of spontaneous speech for most of the languages, only for G1 and Italian read speech was recorded. Due to the high amount of non-natives and dialect speakers who often try to speak the standard language there are a couple of hesitations and corrections.

The size of the vocabulary differs much among the different tasks and languages. The smallest vocabulary size is observed for the train/flight information domain with around thousand words per language. For the other domains, land register data-entry and date scheduling the vocabulary is higher and varies among 2000 and 7000 words depending on the language. For the experiments we tried to limit the recognition vocabulary to a smaller and equal size for all languages in the experiments without language modeling, but left the original size of the lexicon for the experiments with language models.

4. PHONE SUBSTITUTIONS

Each language has its own characteristic set of phonetic units, and from the phones, different phoneme systems may be built. For example, in Japanese, no distinction is made between /r/ and /l/ and they would thus belong to the same phoneme class in that language, whereas in other languages they are phonemes classes on their own since a semantic difference occurs such that words get a new meaning when e. g. /r/ is replaced by /l/. Some sounds are also unique to some languages, for example the vowel /y/ appears within these languages only in German. If recognition is performed for German with a recognizer that was trained with other languages, the sound /y/ must be modeled although it was not represented in the training material. Thus, the parameters of /y/ must be estimated from other vowels like /I/. Sometimes there is the same symbol used for sounds of different languages, but the acoustic properties differ for these sounds. When recognizing multiple languages simultaneously, it may thus be reasonable to share some sounds across languages and to stay with monolingual units for other sounds.

Thus, for both approaches of cross-lingual and simultaneous multilingual recognition, relations and similarities among sounds of different languages must be found.

In general, we can distinguish between a 1:1 mapping of phones between languages and a n:1 or 1:m mapping of phones, which would mean that for example the parameters of /y/ are estimated as e. g. the mean values of /I/ and /u/. In this work we will refer to the first strategy of a 1:1 mapping. In a rough classification, we distinguish among three different approaches within the 1:1 mapping.

na(t)ive approach: this approach follows the principle a non-native would follow when speaking a second language: he basically has the phonetic inventory of the first language and partially uses that inventory when speaking the second language. Some of the new phones can be learnt by a language learner, but they are not always pronounced correctly, and under stress condition or within difficult words a non-native may fall back to his native phonetic inventory. For example Japanese speaking English or German often confuse the use of /r/ and /l/.

phonetic approach: this strategy follows principles in the production of sounds in the human vocal tract. These characteristics for the production of sounds can be classified into place and manner

of production, where the first describes, where obstacles are put in the air flow and which organs are involved in the production of sounds, and the second one describes the manner in which the obstacles act, e. g. a complete or partial closure of the air flow.

Thus, for consonants it can be distinguished with regard to the manner among stop–fricative–approximant–lateral–rhotics and others and for the place between labial–dental–alveolar–palatal-velar–alveolar and others. Another criterion is the voicing of consonants which can be either voiced or unvoiced. For vowels, different tongue positions are distinguished like front–central–back, and for the opening of the mouth among close–close-mid–open-mid–open as well as between rounded and unrounded for the shape of the lips.

The difference between consonants is clearer than between vowels, e. g. a plosive has a complete closure, while others do not have a complete closure, and there is no sound between e. g. a plosive and a fricative. For vowels, the position of the tongue can gradually change and there are transitions between a front and a central vowel, so the distinction and classification of vowels can be more difficult.

For the substitution of sounds in this approach, that sound that agrees in the most phonetic features with the untrained one is taken instead of the unknown one of the goal language. For example, /p/ (plosive, labial, unvoiced) may be replaced by /b/ (plosive, labial, voiced) or by /t/ (plosive, dental, unvoiced). Some hierarchy has to be built in order to define which of the criteria will be changed first.

data-driven approach: this approach determines the similarity among phones with the data given by the trained recognizer. This approach is only possible if there is training data available for the new language, i. e. some adaptation data or for the case of simultaneous multilingual recognition for the decision if acoustic units should be joined.

Measures for the similarity can e. g. be estimated from the Gaussian densities or the codebook parameters of a trained recognizer. Therefore a recognizer must be trained with all languages, and for all observations of a language-dependent sound the similarity parameters like mean values must be estimated and then according to a distance measure the most similar units may be joined. This merging of units can happen in one or more steps and it may also be allowed to split units. The advantage of this approach is that there is no human knowledge or manual work necessary to estimate similarities, but the disadvantage may lie in an exact determination of the segmentation of the speech signal into sounds and consequently an error prone measure for similarities among sounds.

The phonetic description of consonants separates better into classes while measures for the classification of vowels correlate with formant frequencies and of these formant frequencies every compromise between two vowels of, say, 500 and 600 Hertz is possible and thus really different sounds may occur. On the other hand, this characteristic may make it easier to calculate the parameters of sounds by mixing sounds which would average in the same formant frequency.

Another decision is the type of acoustic units that will be used for the target recognizer, especially if the units ought to be mono- or multilingual. For example, to decide for n available languages each containing the sound /a/, if the sound /a/ for the target language (without own training material) shall result from one /a/ of a language or from a mixture of a certain number of /a/'s. With substitution approach 1 and two, the multilingual units may be trained together, and with approach 3 it may be determined according to the data if all or only a couple of /a/'s shall have an influence on the modeling of the new /a/.

Comparing the results of these different strategies for phone substitution it can be found that approaches 1 and 2 are quite similar, of course depending on the priorities set for substitution to manner or place in approach 2. Differences occur mostly when the orthography proposes the pronunciation of another native sound than the similarity according to acoustic features would propose it. For example, in the na(t)ive approach, /u/ may be replaced by /U/ according to the same orthographic spelling [u] rather than to the possibly phonetically closer /o/ if the corresponding criterion is chosen.

Approach 3 is only possible if a certain amount of data is available for all languages; in general it is used for the design of multilingual acoustic units. Errors in this approach can occur if there is not sufficient data available for each language and thus the parameters have not been well estimated. Another source of error for the third approach may be given when the labeling of the speech material according to acoustic units is not completely correct, e. g. with automatic segmentation. Sometimes, silence is assigned to a certain sound and changes this way the statistic properties of this sound.

Another source for errors may be different recording conditions. A consequence may be that sounds of the same language without respect to their phonetic features are estimated as more similar than any sound of the other language. In our experiment, this happened for Slovenian sounds which were for many cases more similar than any sound of another language.

One special phenomenon that has arisen in data-driven decision is the similarity of /j/ and /z/ which have quite different phonetic characteristics (approximant–palatal–voiced vs. fricative–alveolar–voiced) , which has also been shown in several other approaches [5, 6], thus there may be some other measures important besides the phonetic features determined so far.

5. EXPERIMENTS

For our recognition experiments we used the ISADORA recognizer [9] with semi-continuous Hid-

den Markov Models. We performed experiments both with and without language models, for the experiments without language models we used a reduced recognition vocabulary in order to limit the perplexity of the task.

Instead of the technique of polyphones with context-dependent acoustic units we only used monophones with the phone itself and no context around. The performance decreases by using context-free acoustic units, but only with these units we can hold the number of acoustic units and, even more important, the number of necessary substitutions at a relatively low level.

As baseline systems, we ported our recognition system to the new languages and use the performance obtained with monolingual recognizers for our cross-lingual experiments.

Concerning acoustic units, we considered sounds represented by the same phonetic symbol as identical, and thus, for our cross-lingual experiments, we have to replace those phones whose symbol does not occur in the target language. Furthermore, we did not count replacements for the length of phones, i. e. if there existed only a long vowel like /i:/ and the short correspondent /i/ was needed, we did not count this as substitution. The same is done for Italian geminates, thus /nn/ was set equal to /n/ and the substitution was not counted.

In Table 2 the number of substitutions across languages is shown. There are no substitutions between G1 and Italian since they share proper names of both languages and thus phones of both languages are modeled for each recognizer. Between G1 and G2 there are two substitutions for originally Italian phones (/J/, /L/) which are used in the G1 recognizer. There is a high number of substitutions between the Germanic languages (English, German) on the one side and the Slavic languages (Slovak, Slovenian, Czech) on the other side, once due to the high number of consonants modeled in the Slavic languages and the high amount of vowels in the Germanic languages.

Furthermore, we can observe, that, using the Japanese recognizer for the recognition of any of the other languages, a high number of substitutions has to be made, since the phone inventory of the Japanese language is small in comparison to those of the other languages. On the other hand, for recognition of Japanese with any other recognizer, only a small number of substitutions has to be performed.

Furthermore, we have listed in that table also the number of substitutions for multilingual recognizers, and, of course, the number of substitutions decreases with respect to the corresponding monolingual recognizers, although the complete phone inventory cannot be covered with three languages for all others. We have found out that, besides Japanese, that the phone inventory of the remaining 6 languages can only be covered without substitution only when all 6 languages are involved into training, thus there is no real multilingual inventory possible with a subset of these languages.

We performed experiments with na(t)ive and phonetic substitution as well as some preliminary experiments with data-driven substitution for the cross-lingual experiments.

Rec \ Lg	G1	It	Sa	Se	Cz	Jp	En	G2
It	0	0	3	1	4	0	8	0
G1	0	0	3	1	4	0	8	0
Sa	10	10	0	4	6	4	12	11
Se	9	9	5	0	7	2	9	8
Cz	12	12	7	5	0	3	11	11
En	11	11	8	3	7	3	0	9
Jp	12	12	9	6	9	0	13	10
G2	2	2	4	0	5	0	7	0
It-G1	0	0	3	1	4	0	8	0
Se-Sa	7	7	0	0	3	2	8	7
Sa-Se-Cz	7	7	0	0	0	2	8	7
G2-En	2	2	3	0	4	0	0	0
G2-En-Jp	2	2	3	0	4	0	0	0

Table 2. Substitution of phones with different languages and recognizers

6. RESULTS

The experiments performed for this contribution are done without optimization, i. e. without using the technique of polyphones for acoustic units, without using a polygram verification for language modeling and without optimizing the training procedure in order to obtain recognizers trained at the same level. Thus, the results given here, do not correspond to the optimally trained recognizers, but are comparable to each other with respect to modeling and training.

Results of the experiments with language modeling are given in Table 3 for monolingual and cross-lingual recognition, where the monolingual results are shown in the diagonal. We also give some experiments for multilingually trained recognizers in the second part of that table.

Using different strategies for phone substitution did not lead to significant differences between the na(t)ive and the phonetic approach, but often the na(t)ive approach seems lightly better compared to the replacing strategy proposed by [5]. With data-driven substitution, we found substitutions that correspond roughly to phonetic similarities for Italian and G1 data, but for other languages the similarities do not correspond to phonetic properties. For Slovenian, for example, the phones classified as most similar were in most cases also Slovenian phones, probably the recording conditions dominated over the phonetic similarities.

For all languages besides German G2, recognition is best for the monolingual recognizer trained with data of that language and domain. For G2, recognition showed to be better for the bilingual German-English recognizer under these conditions.

The performance among the languages differs from 37 % for G2 to 94 % for Italian. There are various reasons for this difference: the domains have a different difficulty, in the SPEEDATA task the best recognition is achieved, followed by SQEL and finally the VERBMOBIL task. There are different types of speech and other recording conditions with hesita-

Rec \ Lg	It	G1	Sa	Se	Cz	En	Jp	G2
G1	80.96	87.89	28.05	30.96	55.34	8.49	17.89	20.61
It	94.22	70.74	22.19	38.61	59.03	7.44	18.31	18.70
Slovak	77.60	57.07	88.33	71.03	68.91	7.47	20.15	2.59
Slovenian	86.63	60.66	66.60	90.26	52.25	8.87	30.20	2.01
Czech	81.45	57.07	35.02	58.51	88.57	10.54	22.02	5.85
English	41.41	36.14	35.35	26.77	42.71	48.16	20.28	3.07
Japanese	83.30	56.78	40.70	44.11	36.33	5.48	64.53	1.05
G2	81.53	67.59	39.63	59.40	53.49	12.19	25.19	37.10
G1-It	94.14	86.72	28.19	43.22	63.87	8.81	27.29	19.46
Sa-Se	85.83	60.61	84.23	88.00	62.43	7.41	27.99	1.82
Sa-Se-Cz	86.74	65.02	84.05	85.70	83.88	7.16	29.97	1.53
En-G2	84.40	77.13	38.33	60.71	62.28	24.54	29.60	46.98
En-G2-Jp	87.91	73.89	46.37	64.22	64.42	24.10	52.38	46.50

Table 3. Recognition results for cross-lingual experiments

tions, background noise etc. Furthermore, the size of the vocabulary is different for each language. Finally, the languages themselves differ in the difficulty for recognition, some languages may be easier to be recognized than others due to the phonetic structure, word length and other reasons.

In order to compare the performance of the cross-lingual recognizers trained with one language we averaged the performance of all recognizers besides the one of the original language and domain. Best cross-lingual recognition averaged over the seven other recognizers was achieved for Italian with 78.73 %, worst performance was achieved for G2 with 11.37 %. The ranking in the recognition rate remains the same with respect to the monolingual recognition experiments, only Czech moves one step which could be interpreted that Czech is easier to recognize than Slovenian which moved that step down.

Furthermore, we calculated the ratio of the loss of performance by dividing the cross-lingual performance by the monolingual performance and obtain the same ranking. Here, Italian obtains 85.56 % of the recognition, thus the loss of performance when recognizing with other languages is below 15 % on average, while for G2 with 30.65 % only one third of the performance is achieved.

These both calculations are difficult for interpretation since the similarity of languages and thus the recognizability cannot be taken into account, for example we have two German recognizers in the cross-lingual experiments. Assuming a higher similarity among the Slavic languages, the cross-lingual performance should be higher when recognizing with Slavic recognizers for the Slavic languages than for the others. Furthermore, the cross-lingual recognition of Japanese could be worse because there are no languages similar to Japanese used for recognition.

From these numbers, we can observe, that starting with a poor recognition rate for monolingual recognition, the performance for cross-lingual experiments suffers more than for languages and domains where the performance is already higher itself.

Averaging the performance of cross-lingual recognizers on different spoken languages, we find, that, for monolingually trained recognizers, the best cross-lingual performance was achieved by the Slovenian recognizer which lead three times to the best cross-lingual recognition, whereas Czech, English and Japanese never performed best, thus the Slovenian recognizer seems to be best for cross-lingual recognition in this task. The similarity among languages and therefore their reciprocal cross-lingual performance has a high ranking compared to other languages. Only Slovak and Slovenian showed mutually the best performance for cross-lingual recognizers and may therefore be assumed similar for this speech recognition task, although theoretically, Slovak and Czech should be more similar than those two languages.

For other languages, there is no such symmetry observable, even the two German recognizers do not lead to highest reciprocal results: G1 recognizes best G2, but not vice versa. This may be due to different speaking styles, but more probable to the different speakers, since the speakers of G1 speak with a dialect and with a non-native accent, while the G2 speakers are German natives and do not speak with a strong dialect.

With multilingual recognizers, trained with several languages, performance is worse than with the appropriate monolingual recognizer. Having the target language not included into training, the performance is better than with cross-lingual monolingual recognizers. Unfortunately, for those languages which have the highest cross-lingual performance, no multilingual recognizers were trained, thus often the best monolingual cross-lingual recognizers perform better than the best multilingual recognizers trained in these experiments.

Of the available multilingual recognizers, the G2-English-Japanese recognizer performs best for these data, possibly due to a larger variety in the models provided by Japanese in addition to the Germanic languages models.

7. CONCLUSION

In this contribution, we compared the performance of different monolingual recognizers with respect to cross-lingual recognition. We found with our experiments with non-optimized recognizers (only monophones, no polygram verification in the language models, no optimization in the training), that besides the German G2 task, performance is best for monolingual recognizers. The performance of the different

languages differs due to the different difficulty of the task and also due to differing recognizability of the languages.
When monolingual recognition is already bad, cross-lingual performance gets even worse. Thus, for Italian, the average decrease in performance is 15 %, whereas for G2 only one third is recognized with respect to the monolingual recognizer. Cross-lingual performance does not show strong symmetry in the recognition, only Slovak and Slovenian recognize utterances of the other language better than any other language.
When recognizing with multilingual cross-lingual recognizers, performance gets better than with the corresponding monolingual recognizers. Unfortunately, we have not trained all combinations of recognizers, so the combination of the best monolingual cross-lingual recognizers could not always be tested.
Concluding, we found for these languages and domains, that best performance is obtained with monolingual recognizers. For cross-lingual recognition, the choice of the language for training the recognizer is important for the performance. Furthermore, we found that performance increases if training data of more languages are involved and thus both acoustic units are modeled with more variety and more training material as well as more different acoustic units are modeled overall.

REFERENCES

[1] U. Ackermann, F. Brugnara, M. Federico, and H. Niemann. Application of Speech Technology in the Multilingual SpeeData project. In *3rd Crim-Forwiss Workshop*, Montréal, 1996.

[2] J. Barnett, A. Corrada, G. Gao, L. Gillick, Y. Ito, S. Lowe, L. Manganaro, and B. Peskin. Multilingual Speech Recognition at Dragon Systems. In *Proc. Int. Conf. on Spoken Language Processing*, Philadelphia, USA, 1996.

[3] H. Cerf-Danon, S. De Gennaro, M. Feretti, J. Gonzalez, and E. Keppel. TANGORA — a Large Vocabulary Speech Recognition System for Five Languages. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 183–186, Genova, September 1991.

[4] P. Dalsgaard, O. Andersen, and W. Barry. Multi–Lingual label alignment using acoustic–phonetic features derived by neural–network technique. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 197–200, Toronto, Kanada, 1991.

[5] P. Dalsgaard, O. Andersen, and W. Barry. Cross-Language Merged Speech Units And Their Descriptive Phonetic Correlates. In *Proc. Int. Conf. on Spoken Language Processing*, volume 6, pages 2627–2630, Sydney, December 1998.

[6] C.-H. Jo, T. Kawahara, S. Doshita, and M. Dantsuji. Automatic Pronunciation Error Detection And Guidance For Foreign Language Learning. In *Proc. Int. Conf. on Spoken Language Processing*, volume 6, pages 2639–2942, Sydney, December 1998.

[7] J. Köhler. Multi-lingual Phoneme Recognition Exploiting Acoustic-phonetic Similarities of Sounds. In *Proc. ICSLP'96*, Philadelphia, USA, 1996.

[8] E. Nöth, S. Harbeck, H. Niemann, V. Warnke, and I. Ipšić. Language Identification in the Context of Automatic Speech Understanding. In N. Pavesic, H. Niemann, S. Kovacic, and F. Mihelic, editors, *Speech and Image Understanding*, pages 59–68. IEEE Slovenia Section, Ljubljana, Slovenia, 1996.

[9] E. G. Schukat-Talamazzini. *Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen.* Künstliche Intelligenz. Vieweg, Braunschweig, 1995.

[10] F. Weng, H. Bratt, L. Neumeyer, and A. Stolcke. A Study of Multilingual Speech Recognition. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 359–362, Greece, September 1997.

[11] S. Young, M. Adda-Decker, X. Aubert, C. Dugast, J.-L. Gauvain, D. Kershaw, L. Lamel, D. Leeuwen, D. Pye, A. Robinson, H. Steeneken, and P. Woodland. Multilingual large vocabulary speech recognition: the European SQUALE project. *Computer Speech & Language*, 11:73–89, 1997.